CS 591 Project Report: Analyzing Bike Rack Data in Cambridge

By Jacob Levy and Cameron Sonn

Introduction:

Our project seeks to improve environmental conditions in the greater Boston and Cambridge areas. After searching different open data portals, our initial approach focused on examining open space and public biking data sets. We performed basic analysis of open spaces in both areas by gathering data on how much open space there was in total and looking at the standard deviation and mean value of the open spaces in order to get a sense of the sizes of spaces we were working with. Although this in itself was a success, we could not determine how we could use this data to make the greater Boston and Cambridge areas more environmentally friendly. We decided to then turn to our data on bike paths and bike racks in the Cambridge area. We decided that adding in or adjusting the location of the existing bike racks in this area could have a large impact on the environment around it since more people in remote locations that previously didn't have access to bike racks now would. We decided to focus on the placement of new bike racks for the remainder of our project and to treat it as an optimization problem.

Methods:

We pulled datasets from three different sources: Analyze Boston, Cambridge Open Data, and Cambridge GIS data from GitHub. From Analyze Boston we used their GeoJSON API to pull Boston Open Space data and Bike Path Data. From Cambridge Open Data we used their GeoJSON API to pull Cambridge Open Space data. From GitHub we used GeoJSON files to pull Cambridge Bike Rack data and Bike Path data. We collected these data sets and stored them in a local MongoDB repository.

Once the five datasets were collected, we created three new datasets using the relational paradigm. For our first data set, we combined the open space data sets from Boston and Cambridge by projecting out the areas of each open space. The GeoJSON files from each city were not structured the same for each city so we needed to manually inspect each file to correctly index the areas of each collection of open spaces. We then used a union to join the two sets together into one new dataset. For our second dataset we used a similar approach. We created a

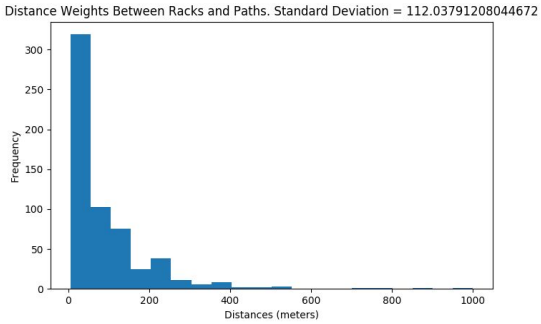
new joint bike path dataset by projecting out the lengths of each bike path. We then joined the two sets together with a union to create a joint bike path data set. For our third data set we looked at the Cambridge Bike Path data and Cambridge Bike Rack data. We projected out the coordinates of the bike paths and bike racks and joined them into a third new dataset. We needed to perform some manual operation to fix the format of the Cambridge bike path coordinate data as some of the GeoJSON linestring coordinates were not properly formatted most likely due to human error.

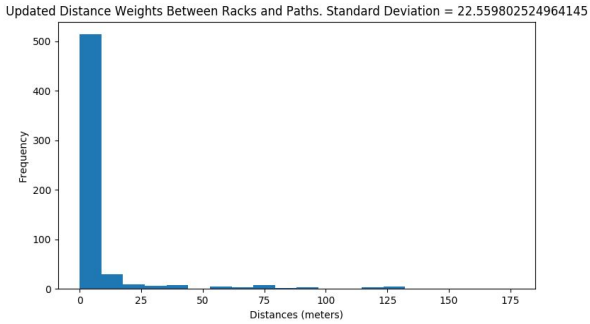
In the first project, we performed simple calculations such as comparing the ratio of bike racks to bike paths, determining the total number of bike racks, and determining the total length of all the accumulated bike paths. With this new data and the new datasets we had made for ourselves we were better able to tackle the problem we had set out to solve.

As an arbitrary decision, we decided to add five hundred new bike racks to the city of Cambridge. We determined that the areas that would make the most sense to add these new bike racks would be the start or end points of the different bike paths in the locations furthest removed from preexisting bike racks. In order to find these locations we calculated the distances from each potential point for a new bike rack to every other already existing bike rack. This allowed us to form a list of the longest distances, or the locations that were most in need of new bike racks. After performing other calculations to help support our findings, we were confident that adding bike racks to these new locations would help increase the amount of people in the Cambridge area that would have access to a more environmentally friendly means of transportation.

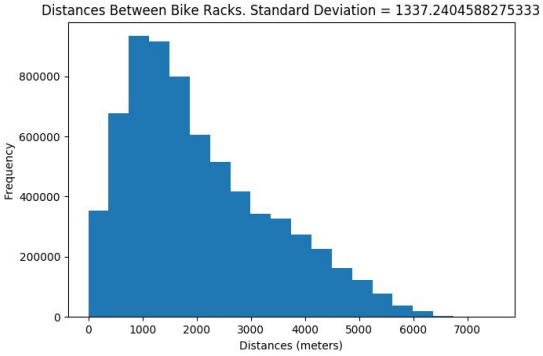
Results:

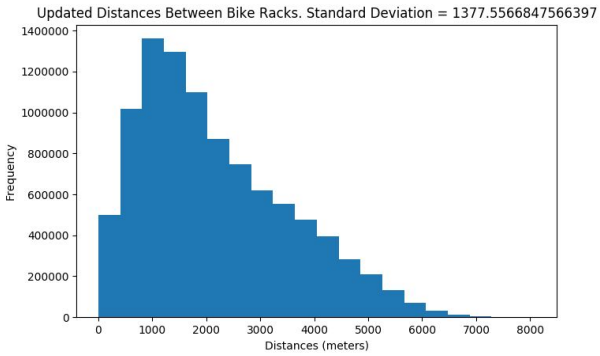
In order to get a sense of how successful we were in our optimization problem, we found the standard deviation of the closest distances to the nearest bike racks for our data set before and after we inserted the extra coordinates we had determined to be most useful to solving our problem. We made histograms for each of these sets.





The updated histogram was skewed towards the shorter distances, which made sense because the potential spots for bike racks will find themselves and return 0.0 as their shortest distance since we added them in already. We used this to prove to ourselves that we had added in the coordinates correctly. Next we compared all the distances from each bike rack to every other bike rack before and after we added in the new ones and made new histograms and found their standard deviations.

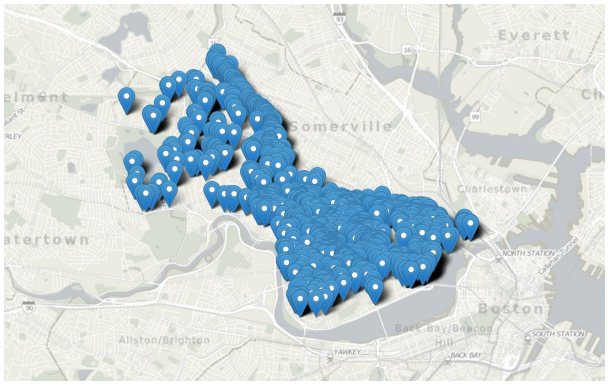




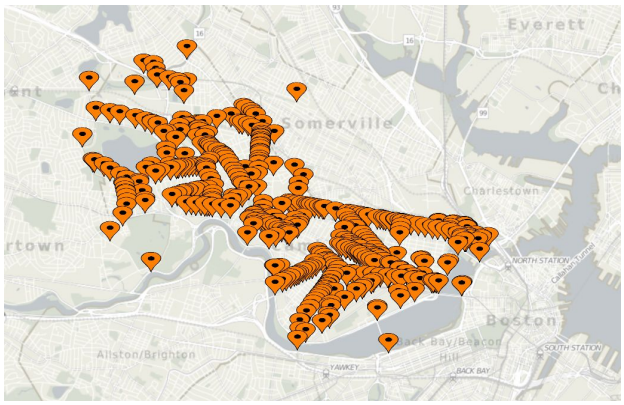
The histograms are quite similar but the range of lengths in the updated one extends slightly further out to the right which shows that we added in bike racks that were further away from all the preexisting bike racks. The standard deviations were also very similar with the updated one being slightly higher. Overall we believe that we successfully added in new bike racks in a way that fit our goal, and showed statistically that we calculated them correctly.

Once we finished our calculations, we chose to visualize our new bike rack data with Leaflet.js. We used Leaflet in conjunction with an HTML document to display our different data sets. Our web page supports three different visualizations.

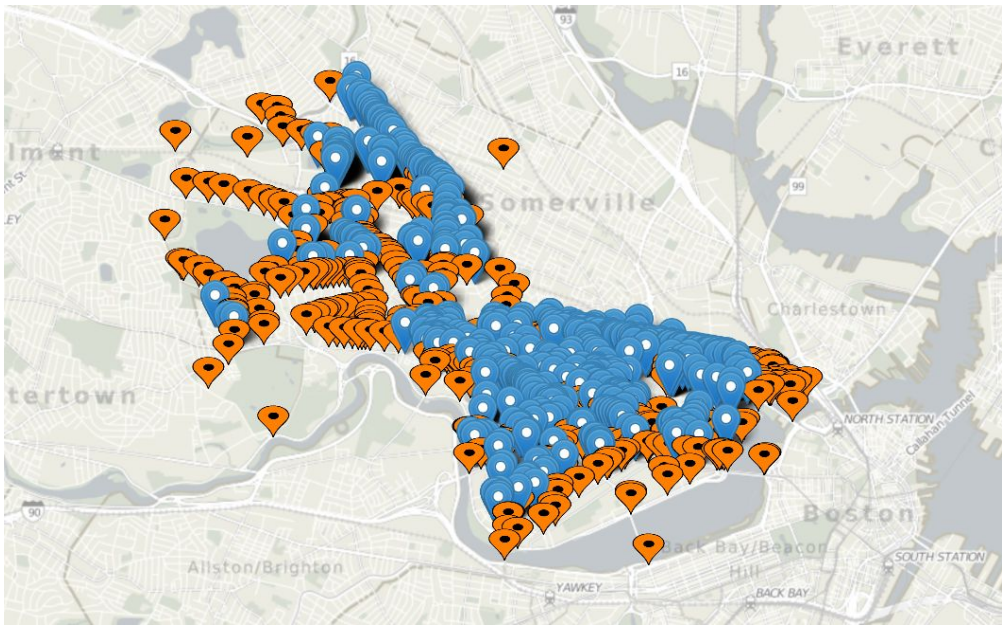
The original bike rack locations:



The five hundred new rack locations:



The new and original rack locations combined together in one map:



Future Work:

In order to attain more accurate readings on where to add in the additional bike racks to fit this optimization problem, future work may include factors such as population density. For instance, our current calculations are based off of distances from potential bike rack locations to currently existing ones in order to provide bike racks to areas relatively far away from existing clusters. However, one question arises: what if one of the areas where we decided to add the bike racks was indeed the furthest away, and therefore the highest priority, but also had a much lower population density than an area somewhat closer to pre existing clusters? If our goal was to maximize the amount of people in different locations to have access to biking as a form of transportation, it would make the most sense to also target highly populated areas. One of the reasons we did not focus on this in our original project was that this had already been done for us in some part. Take for example Harvard Square. This is a highly populated and visited area and it contains one of the largest clusters of currently existing bike racks. We reasoned that the people who originally installed the bike racks would have taken population density into account and so we made it a lower priority. However, population densities change over the years and future work in this area would help to strengthen our findings.

Another extension of our work that we could do in the future is to apply it to areas outside of the Cambridge area. If our numbers and calculations were correct for the subset that is the Cambridge area, then they could also potentially be applied to the greater Boston area, or even other locations in the world. Finally, we could then analyze the environmental impact of our new racks and determine if their presence made a significant difference.